图书情报工作
第64卷 第11期 2020年6月

《史记》历史事件自动抽取与事理图谱构建研究*

■ 刘忠宝[1,2] 党建飞[2] 张志剑[2]
[1] 云计算与物联网技术福建省高等学校重点实验室(泉州信息工程学院) 泉州 362000
[2] 中北大学软件学院 太原 030051

摘 要:[目的/意义]《史记》是我国第一部纪传体史书,几乎囊括黄帝时代到汉武帝元狩元年3 000多年的重大历史事件。如何快速准确地发现这些历史事件及其之间的内在联系,对于透过历史现象、揭示历史实质以及发现历史规律具有重要意义。[方法/过程]在BERT模型和LSTM-CRF模型的基础上,提出面向《史记》的历史事件及其组成元素抽取方法,并基于此构建《史记》事理图谱。[结果/结论]实验结果表明,利用所提方法抽取历史事件及其组成元素的F1值分别达到0.823和0.760。通过事理图谱能够发现蕴含在《史记》中鲜为人知的知识,这为文献学、历史学、社会学等领域专家开展研究提供必要的资料准备。

关键词:《史记》 历史事件抽取 事理图谱 BERT模型 双向长短期记忆网络 条件随机场
分类号:G256
DOI:10.13266/j.issn.0252-3116.2020.11.013

1 引言

习近平总书记在党的十九大报告中指出,文化是一个国家、一个民族的灵魂。文化是一个国家和民族精神的延续,而优秀的传统文化是一个国家和民族文化与精神层面的集中表达。从历史中汲取知识、获取经验,并将其转化为解决前进道路上种种问题和重重困难的制胜法宝,是实现中华民族伟大复兴的不竭动力和力量源泉。作为优秀传统文化载体的中华典籍,在漫长的历史发展历程中不断丰富发展,最终形成了具有中华民族特色的文化宝藏。在众多中华典籍中,《史记》一直占据着重要地位,不仅因为它是我国纪传体史学的奠基之作,也是我国传记文学的开端,至今仍被世人推崇。

《史记》共有130篇,50余万字,记载了自上古传说中的黄帝时代,到汉武帝元狩元年间共3 000多年的历史。其每一个历史人物和每一起历史事件都是对史实反复核准后写就的,对于如何在这部恢弘巨著中快速准确地发现历史事件及其之间的内在联系,进而透过历史现象,揭示历史实质,发现历史规律具有重要意义。历史事件之间在时间和空间两个维度上的演化过程和规律具有重要的研究价值。当前广受关注的知识图谱着力构建实体及其关系的知识网络,缺乏对历史事件及其关系的刻画。事理图谱的出现能够有效地弥补上述不足。在组织结构上,事理图谱是一个有向图,其中节点表示历史事件,节点的属性表示历史事件的组成元素,有向边表示历史事件之间的关系。

基于上述分析,笔者面向《史记》语料集,在BERT模型(Bidirectional Encoder Representations from Transformers)和LSTM-CRF模型的基础上,提出《史记》历史事件及其组成元素抽取方法,并基于此构建《史记》事理图谱,以揭示历史事件的发展过程和演化规律,全面刻画历史人物的行为活动,为文献学、历史学、社会学等领域专家开展研究提供了必要的资料准备。

2 研究进展

事件抽取是从非结构化的语料集中自动抽取事件的信息并以结构化的方式表示。事件抽取是事理图谱构建的关键。事件抽取方法包括基于模板匹配的方法、基于机器学习的方法以及基于深度学习的方法3类。

基于模板匹配的方法的基本思路是利用人工标注的语料集来进行事件的抽取。根据人工参与度的多少,该方法可分为有监督的方法和弱监督的方法。有

* 本文系国家社会科学基金一般项目“大数据环境下面向图书馆资源的跨媒体知识服务研究”(项目编号:19BTQ012)研究成果之一。
作者简介:刘忠宝(ORCID:0000-0002-0038-2462),教授,博士,E-mail:liu_zhongbao@hotmail.com;党建飞(ORCID:0000-0002-7419-0455),硕士研究生;张志剑(ORCID:0000-0002-7758-9277),硕士研究生。
收稿日期:2019-12-04 修回日期:2020-02-06 本文起止页码:116-124 本文责任编辑:徐健

刘忠宝，党建飞，张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J]. 图书情报工作,2020,64(11):116 - 124.

监督的方法根据人工标注的语料集进行事件抽取。E. Riloff 等在构建触发词词典的基础上,综合利用事件元素的描述信息及其之间的上下文语义关系,构建了 13 种事件匹配模板[1];J. T. Kim 等基于 WordNet 词典,利用短语结构和语义框架,构建了并行化的事件匹配模型 PALKA[2]。弱监督的方法只需对部分语料集进行人工标注即可进行事件抽取。E. Riloff 等并未标注语料集中的所有事件元素,只标注了事件类型,便可基于预分类语料进行事件抽取[3],该研究极大地减少了语料集的人工标注量;姜吉发提出一种基于领域无关概念层次知识库的事件模式学习方法,该方法无需人工标注语料集以及事件类别,只需给出事件抽取任务定义就能完成原始语料的事件抽取,该方法将事件抽取模式划分为语义模式、触发模式、抽取模式、特例模式和泛化模式等[4];许君宁等利用 HowNet 的语义角色框架标注事件的语义角色,进而完成从非结构化文本中抽取事件的任务[5]。基于模板匹配的方法对特定领域的事件抽取表现优异,但事件模板需要大量人工标注,耗时耗力,且事件模板存在适应性差的问题,无法解决通用领域的事件抽取问题。

基于机器学习的方法将事件抽取问题转化为分类问题,进而利用机器学习算法进行事件抽取。基于机器学习的方法分为 4 个阶段[6]:首先,判断词语是否是触发词和事件类型;然后,判断词语是否是事件元素;接着,判断事件属性;最后,进行事件共指消解。常用于事件抽取的机器学习模型有支持向量机、最大熵、隐马尔科夫模型等。S. Saha 等利用支持向量机模型进行分子生物事件提取[7];F. Zhu 等利用最大熵模型提取中文事件[8];许旭阳等提出基于事件实例驱动的新闻文本事件抽取方法,该方法首先抽取事件特征以形成候选事件实例的表示,然后利用分类算法来判断是否是事件实例,最后利用层次聚类算法 k-medoids 进行事件抽取[9];刘振利用条件随机场模型和语义角色标注技术,提出面向网络科技信息的事件抽取方法[10];吉久明等对支持向量机、条件随机场、聚类算法等中文事件抽取算法进行了比较研究,研究结果表明,除应用条件随机场抽取个人简历类格式规范的语料中事件取得优异的效果外,F 值普遍低于 0.9[11]。基于机器学习的方法在很大程度上减少了人力投入,其适应性和工作效率有了较大提升。然而,该方法的各阶段相对独立,前面阶段的误差很可能传递到后面阶段,事件抽取性能也随之降低。

随着深度学习的广泛应用,深度学习逐渐成为事件抽取的主要方法。常用于事件抽取的深度学习模型有卷积神经网络、递归神经网络、长短期记忆网络等。Y. Chen 等[12]和 T. Nguyen 等[13]最先将卷积神经网络模型引入到事件抽取中。前者基于卷积神经网络模型,引入动态多池机制来提高事件抽取效率。该机制根据位置信息将候选触发词和候选实体进行分割,较之最大池机制,能够获得更深层次的特征信息[12]。后者提出一种基于 Skip-gram 的卷积神经网络模型,该模型能够高效地提取非连续短语的特征,因而能够高效地完成事件抽取任务[13]。递归神经网络模型擅长处理序列化的语料信息。因此,X. C. Feng 等首先利用递归神经网络模型对语料中的每条句子进行序列建模,进而获得句子的上下文信息;然后,利用卷积神经网络模型获取短语的特征信息;最后,在融合上述两类特征的基础上进行事件抽取[14]。T. H. Nguyen 等提出基于递归神经网络的事件抽取模型。该模型分别利用文本序列和记忆网络发现事件的局部特征和全局特征,在融合上述两类特征的基础上进行事件抽取[15]。目前很多事件抽取研究都是面向句子级的,但 Y. Zhao 等提出在事件抽取时融入文档级的特征,真实数据集上的实验表明,文档级的特征对于提高事件抽取效率发挥了重要作用[16]。此外,基于深度学习的方法往往易受虚假特征信息的影响。鉴于此,Y. Hong 等利用对抗神经网络降低虚假特征信息的干扰,进而提高了模型的学习效率[17]。

上述方法的研究对象是英文语料集或是现代汉语语料集,而面向中华典籍的事件抽取研究还不多见。与此同时,深度学习研究不断深入,2018 年底由 Google 提出的 BERT 模型创造了自然语言处理领域的多项记录[18]。基于上述分析,笔者基于 BERT 模型和 LSTM-CRF 模型,对《史记》历史事件及其组成元素抽取和事理图谱构建问题展开深入研究。

3 数据来源与研究框架

3.1 数据来源

笔者从古诗文网(https://www.gushiwen.org/)爬取《史记》语料集。《史记》由本纪、表、书、世家以及列传 5 部分组成,其中本纪与列传所占的篇幅最多,本纪以时间线为主线记载了各朝代帝王的史事;世家以年系事,记载了王侯封国的历史变迁;列传记载了重要人物的主要事迹;书是有关典章制度的专篇;表以表格的形式记载了历史人物和事件。《史记》共有 130 篇,其中表 10 篇,笔者选取除表之外的 120 篇文献作为实验

语料集。

3.2 研究框架

图 1 给出了《史记》历史事件及其组成元素抽取框架。首先,利用预训练语言模型 BERT 对实验语料集进行向量化表示;接着,根据触发词表,得到语料集中历史事件之间的关系;然后,利用双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)抽取实验语料集的上下文语义特征,得到与事件关系相关的候选历史事件,利用条件随机场(Conditional Random Field, CRF)的约束规则确定最终的历史事件;最后,利用 BiLSTM-CRF 从历史事件中抽取其组成元素。以历史事件为节点,历史事件组成元素为属性,历史事件关系为边,构建《史记》事理图谱,并利用图数据库 Neo4j 对事理图谱进行存储。

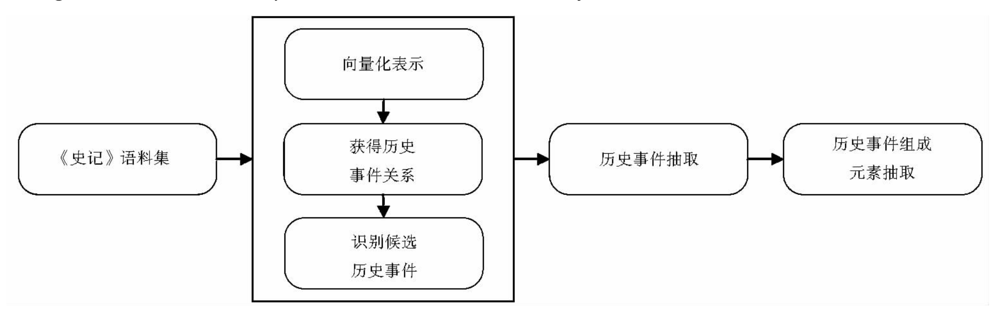


图 1　历史事件及其组成元素抽取框架

4 模型引入

历史事件及其组成元素抽取均采用 BERT 和 BiLSTM-CRF 混合模型。以历史事件抽取为例,介绍两类模型的基本工作流程。

4.1 BERT 模型

BERT 模型利用 Transformer 双向编码表示,通过引入自注意力机制,能够更好地描述历史事件上下文的语义特征。该模型有效地解决了传统向量表示方法由于对历史信息过分依赖而出现的"一词多义"的问题,图 2 给出了 BERT 模型的整体结构。利用 BERT 模型对实验语料集进行向量化表示的基本流程具体如下:首先,依次将语料集中的每个句子输入模型;接着,将输入的句子表示为由字向量、句向量和位置向量组成的输入向量 $E_i(i=1,2,\cdots,n)$;然后,利用多层 Transformer(图 2 中简写为 Trm)对语料经特征提取后生成特征向量 $T_i(i=1,2,\cdots,n)$。

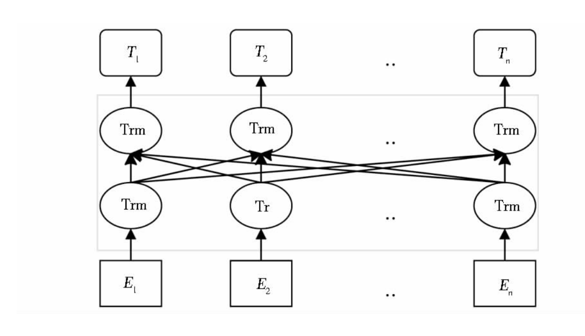


图 2　BERT 模型结构

4.2 BiLSTM-CRF 模型

长短期记忆网络模型(Long Short-Term Memory, LSTM)通过引入门结构来决定在训练期间需要保留或遗忘的信息[19],因而该模型适用于处理长序文本。LSTM 由输入门 i_t、输出门 o_t 和遗忘门 f_t 以及记忆单元 c_t 组成。其中,遗忘门用来控制历史信息,输入门用来控制当前信息,输出门用来确定下一个隐藏层状态,记忆单元用来保存历史信息。LSTM 的工作流程由以下公式表征:

$$i_t = sigmoid(W_i * [h_{t-1}, x_t] + b_i) \quad 公式(1)$$

$$f_t = sigmoid(W_f * [h_{t-1}, x_t] + b_f) \quad 公式(2)$$

$$o_t = sigmoid(W_o * [h_{t-1}, x_t] + b_o) \quad 公式(3)$$

$$c_t = f_t * c_{t-1} + i_t * tanh(W_c * [h_{t-1}, x_t] + b_c) \quad 公式(4)$$

$$h_t = o_t * tanh(c_t) \quad 公式(5)$$

其中,$sigmoid$ 和 $tanh$ 函数为激活函数,x_t 表示 t 时刻的输入,h_t 表示隐藏层单元,W_i、W_f、W_o、W_c 和 b_i、b_f、b_o、b_c 分别表示对应的权重矩阵和偏置。

实验语料集可以看作是一种长序文本,该语料集中的上下文具有紧密的联系。然而,传统的 LSTM 模型只能利用 t 时刻之前的信息,而无法利用 t 时刻之后的信息。因此,笔者引入双向长短时记忆神经网络 BiLSTM 模型[20],该模型由两个方向相反的 LSTM 组成,这种结构能够充分利用语料集的上下文进行历史事件抽取。

图 3 给出 BiLSTM-CRF 模型的整体结构。利用 BiLSTM-CRF 模型对实验语料集进行历史事件抽取的

基本流程具体如下：首先，将 BERT 模型得到的特征向量 $T_i(i=1,2,\cdots,n)$ 输入模型；然后，利用 BiLSTM 模型学习特征向量之间的语义关系，并为每个特征向量打上相应的历史事件标签；最后，根据 CRF 的约束规则分析历史事件关系标签之间的语义关系，进而得到历史事件抽取结果。

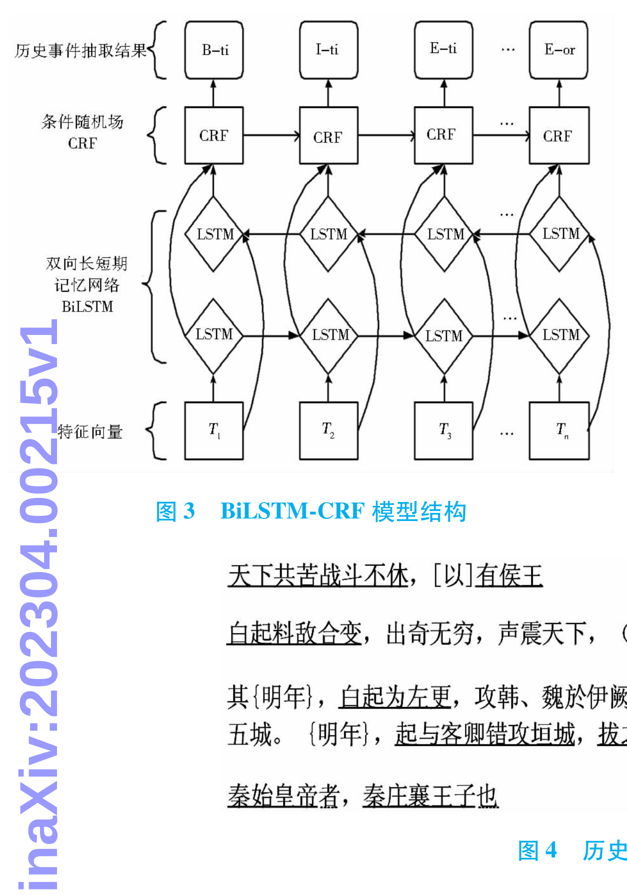


图 3　BiLSTM-CRF 模型结构

5　实验设计和实验结果分析

5.1　语料预处理

历史事件抽取包括两类任务：一类是抽取历史事件，另一类是抽取历史事件组成元素。

在抽取历史事件时，根据历史事件的关系，结合实验语料集的语句结构，给出历史事件关系的触发词，如表 1 所示：

表 1　历史事件关系及其对应的触发词

历史事件关系	含义	触发词
并列	两件事件同时发生	着..也、同年等
转折	某事件与下一事件发生反转	然、以为等
顺承	某事件接着一件事件发生	其年、明年等
因果	两件事件构成因果关系	以、乃等

根据表 1 所示的历史事件关系及其对应的触发词，对历史事件进行人工标注，标注结果如图 4 所示，其中标有“____”符号的语料表示历史事件，标有其他符号的语料表示不同的历史事件触发词。

天下共苦战斗不休，[以]有侯王

白起料敌合变，出奇无穷，声震天下，（然）不能救患於应侯

其{明年}，白起为左更，攻韩、魏於伊阙，斩首二十四万，又虏其将公孙喜，拔五城。{明年}，起与客卿错攻垣城，拔之。

秦始皇帝者，秦庄襄王子也

··· 并列
[] 因果
() 转折
{} 顺承
— 历史事件

图 4　历史事件标注样例

在抽取历史事件组成元素时，根据实验语料集的特点，将历史事件组成元素分为时间、地点、参与者 3 部分。以图 4 中部分语料“其明年，白起为左更，攻韩、魏於伊阙，斩首二十四万，又虏其将公孙喜，拔五城。明年，起与客卿错攻垣城，拔之。”为例，进行历史事件组成要素标注，标注结果如图 5 所示：

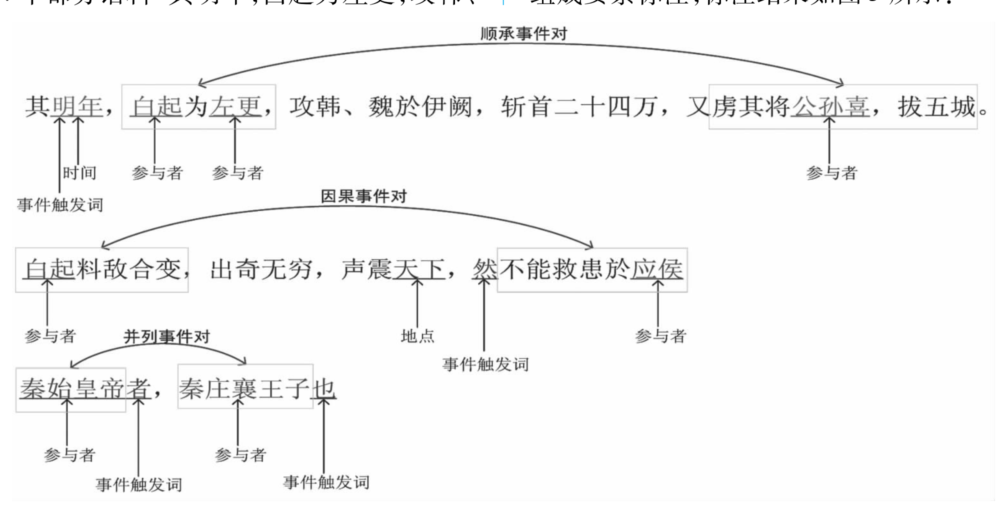


图 5　历史事件组成元素标注样例

为了便于模型训练，笔者定义了 12 类标记对历史事件及其组成元素对应的描述性文本进行标注。其中，“并列”“转折”“顺承”“因果”等历史事件关系分别用{ti,tu,or,ca}标记进行表示，“时间”“地点”“参与者”等组成元素分别用{tm,loc,per }进行表示。历史事件及其组成元素对应的描述性文本分别用{B－,I－,E－}前缀表示历史事件及其组成元素的初始文字、中间文字和结束文字。例如，因果关系事件“天下共苦战斗不休，以有侯王”经序列化标注后，可以表示为“天/B-ca，下/I-ca，共/I-ca，苦/I-ca，战/I-ca，斗/I-ca，不/I-ca，休/E-ca，以有/B-ca 侯/I-ca 王/E-ca”。历史事件及其组成元素对应的描述性文本标记含义分别如表 2 和表 3 所示：

表 2　历史事件对应的描述性文本标记含义

标记	标记的含义
B-ti	并列开始文字
I-ti	并列中间文字
E-ti	并列结束文字
B-tu	转折开始文字
I-tu	转折中间文字
E-tu	转折结束文字
B-or	顺承开始文字
I-or	顺承中间文字
E-or	顺承结束文字
B-ca	因果开始文字
I-ca	因果中间文字
E-ca	因果结束文字

表 3　历史事件组成元素对应的标记含义

标记	标记的含义
B-tm	时间开始文字
I-tm	时间中间文字
E-tm	时间结束文字
B-loc	地点开始文字
I- loc	地点中间文字
E- loc	地点结束文字
B-per	参与者开始文字
I-per	参与者中间文字
E-per	参与者结束文字

5.2　历史事件抽取实验

5.2.1　实验参数设置

为了验证笔者提出的模型 BERT＋BiLSTM-CRF 的有效性，笔者设计了 7 组对比实验，分别是：TF-IDF＋RNN、TF-IDF＋LSTM、TF-IDF＋BiLSTM-CRF、Word2Vec＋RNN、Word2Vec＋LSTM、Word2Vec＋BiLSTM-CRF、BERT＋RNN。随机选取 30 篇实验语料集进行预训练，进而得到最优实验参数。RNN、LSTM、BiLSTM-CRF 等模型，实验迭代次数 epoch 设为 200；为了防止过拟合，将 dropout 设置为 0.5；批量(BatchSize)的大小在网格[16,32,64,128,200,256]中选取。实验性能的评价指标包括准确率 P、召回率 R 以及调和平均值 F1 值，具体定义如下：

$$P=\frac{TP}{TP+FP}\qquad \text{公式(6)}$$

$$R=\frac{TP}{TP+FN}\qquad \text{公式(7)}$$

$$F1=\frac{2P\cdot R}{P+R}\qquad \text{公式(8)}$$

其中，TP 表示正确识别的事件数，FP 表示错误识别的事件数，FN 表示无法识别的事件数。

图 6 给出了上述 8 类模型的批量值与 $F1$ 值的关系。

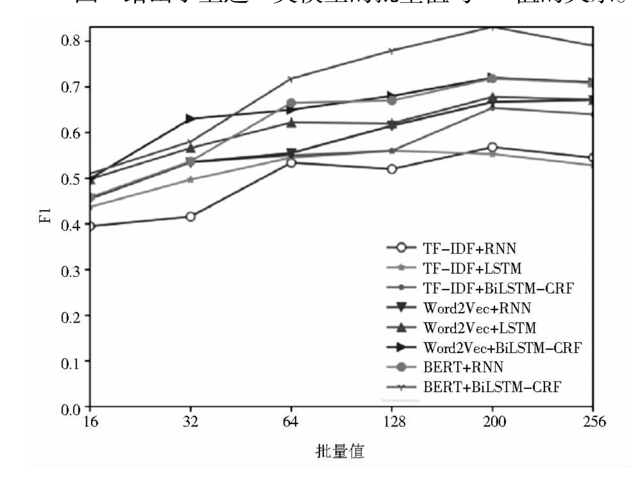


图 6　批量值与 F1 值的关系

由图 6 可以看出，F1 值随批量值变化的趋势是：起初 F1 值随批量值的增大而增大，在达到峰值后，F1 值随批量值的增大而减小。出现这种现象的原因是当批量值较小时，由于模型学习到的语义特征较少，导致模型的表现较差；当批量值过大时，模型的更新周期变长，导致模型的语义特征学习能力变差，F1 值降低。表 4 给出了实验中各类模型的批量值：

表 4　批量值设置表

实验模型	批量值
TFIDF＋RNN	256
TFIDF＋LSTM	128
TFIDF＋BiLSTM-CRF	128
Word2Vec＋RNN	256
Word2Vec ＋LSTM	200
Word2Vec ＋BiLSTM-CRF	128
BERT＋RNN	128
BERT＋BiLSTM-CRF	128

5.2.3 比较实验

为了避免单次实验的偶然性，引入 10 折交叉验证法分别进行 10 组训练和测试，取其中最优的实验结果作为最终实验结果。分别在实验语料集上运行上述 8 类模型，并引入准确率 P、召回率 R、$F1$ 值对模型性能进行评价。实验结果如表 5 所示：

表 5　历史事件抽取比较实验结果

序号	实验模型	*P*	*R*	*F1*
1	TF-IDF + RNN	0.627	0.614	0.620
2	TF-IDF + LSTM	0.564	0.627	0.594
3	TF-IDF + BiLSTM-CRF	0.681	0.683	0.682
4	Word2Vec + RNN	0.723	0.714	0.718
5	Word2Vec + LSTM	0.715	0.738	0.727
6	Word2Vec + BiLSTM-CRF	0.762	0.759	0.760
7	BERT + RNN	0.756	0.747	0.751
8	BERT + BiLSTM-CRF	0.825	0.821	0.823

由表 5 可以看出，在利用 TF-IDF 进行语料向量化表示的实验中，与 RNN 和 LSTM 相比，BiLSTM-CRF 具有最优性能，其 F1 值达到了 0.682。主要原因是历史事件的上下文语义联系密切，RNN 和 LSTM 只能学习到历史事件的“上文”信息，无法利用其“下文”信息，因而性能不高；BiLSTM-CRF 不仅可以从正、反两个方向提取语料中各句的语义特征，而且通过引入 CRF 约束规则可以给出更为准确的抽取结果。在分别利用 Word2Vec 和 BERT 模型进行语料向量化表示的实验中，上述结论同样成立。对比 TF-IDF、Word2Vec、BERT 三种向量化表示模型的实验结果不难看出，基于 TF-IDF 的历史事件抽取 F1 值最高达到 0.682，基于 Word2Vec 的历史事件抽取 F1 值最高达到 0.760，基于 BERT 模型的历史事件抽取 F1 值最高达到 0.823，其主要原因是，与 TF-IDF 相比，Word2Vec 能够充分利用特征向量之间的语义关系；与 TF-IDF 和 Word2Vec 相比，BERT 模型具有最优性能，原因在于该模型在向量化表示过程中充分利用了语料集的上下文语义特征。在同一向量化表示模型下，BiLSTM-CRF 均具有最优性能，特别是在 BERT 模型进行向量化表示的前提下，该模型的 F1 值达到了 0.823。由此可见，笔者提出的模型 BERT + BiLSTM-CRF 非常适用于历史事件抽取任务。

5.3　历史事件组成元素抽取实验

利用 BiLSTM-CRF 模型抽取历史事件的组成元素。实验参数与 5.2.1 节中相同。实验结果如表 6 所示：

表 6　历史事件组成元素抽取实验结果

组成元素	*P*	*R*	*F1*
时间	0.795	0.788	0.791
地点	0.724	0.755	0.739
参与者	0.753	0.749	0.751
均值	0.757	0.764	0.760

由表 6 可以看出，BiLSTM-CRF 模型在抽取时间、地点、参与者等历史事件组成元素时表现良好，其准确率、召回率、F1 值均在 0.75 以上。这表明，BiLSTM-CRF 模型能够较好地完成历史事件组成元素抽取任务。

5.4　可视化实验

笔者在 BERT + BiLSTM-CRF 模型的基础上，利用 Python 编程语言开发了面向《史记》的事理图谱可视化系统平台。该平台将历史事件及其关系存储到 Neo4j 数据库，利用 Python 的 Django 框架进行系统的前后台连接。笔者以《商君列传》为例，展示了历史事件的抽取以及事理图谱的构建过程。

图 7 给出了历史事件抽取页面。将《商君列传》语料输入系统，在页面的左侧给出待识别的语料，系统自动调用训练好的模型文件，抽取出该语料包含的历史事件，并展现在页面右侧。该例包含的因果关系事件有 <圣人苟可以彊国，因果，不法其故>、<太子犯法，因果，法之不行，自上犯之>、<今君之见秦王，因果，嬖人景监以为主>，并列关系事件有 <苟可以利民，并列，不循其礼>、<商君相秦十年，并列，宗室贵戚多怨望>、<夫五羖大夫，并列，荆之鄙人>、<缪公知之，并列，举之牛口之下>、<今君又左建外易，并列，非所以为教>，顺承关系事件有 <公孙鞅闻秦孝公下令，顺承，求见孝公>、<孝公用卫鞅，顺承，鞅欲变法>、<卫鞅为左庶长，顺承，卒定变法之令>、<太子不可施刑，刑其傅其师，顺承，秦人皆趋令>、<赵良见商君，顺承，商君弗从>、<秦孝公卒，顺承，太子立>、<公子虔之徒告商君欲反，顺承，发吏捕商君>。为了便于理解，系统给出了历史事件的译文，见图 8。

图 9 给出了事理图谱的可视化界面。该图所示的事理图谱表达的史实是：商鞅听闻秦孝公下令全国寻求有才之人，商鞅因此托景监求见孝公，以获得赏识；而后“鞅欲变法”，引起革新与守旧两派之间的斗争；“卒定变法之令”，开始制定新法的内容；“于是太子犯法”，刑黥太子师傅，以此树立新法的威严；“赵良见商君”指出，商鞅身为国相不为民造福而自行其是，但商鞅并未正视赵良谏言。商鞅在秦国任相十年，虽然百



图 7 历史事件抽取页面



图 8 历史事件译文页面

姓家家富裕充足，但秦国的皇亲国戚一直因太子之事怨恨商鞅，因此在秦孝公死后，太子继位，商鞅被处以刑罚。通过上述事理图谱可以直观地看出：从商鞅通过景监得到秦孝公的赏识，到天下人觉得其"名不配位"，再到其终未采纳赵良谏言，商鞅受刑并非偶然。

6 总结

《史记》是一部纪传体巨著，如何从中挖掘出一些重要的历史事件及其之间的内在联系，对于从事文献学、历史学和社会学等学科的研究具有重要意义。随着大数据时代的到来以及人工智能技术的发展，深度学习模型层出不穷，特别是 2018 年底出现的 BERT 模型，由于其具有强大的特征提取能力而备受推崇。笔者在《史记》语料集的基础上，融合 BERT 模型和 LSTM-CRF 模型，对历史事件及其组成元素抽取以及事理图谱构建方法进行研究。实验结果表明，利用笔者所提方法抽取历

刘忠宝，党建飞，张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J]. 图书情报工作,2020,64(11):116－124.



图 9　事理图谱展示页面

史事件及其组成元素的 F1 值分别达到0.823和0.760。然而，研究仍存在局限性，如语料集的标注采用人工方式进行，该方式费时费力，能否引入智能化技术进行自动标注值得深入研究；笔者只针对单一著作展开研究，能够发现的重要规律和有价值的知识相对有限，能否融合多部著作进行跨文本研究有待于进一步探讨。

参考文献：

[1] RILOFF E. Automatically constructing a dictionary for information extraction tasks[C]//Proceedings of the 11th national conference on artificial intelligence. Menlo Park: AAAI, 1993: 811－816.

[2] KIM J T, MOLDOVAN D I. Acquisition of linguistic patterns for knowledge-based information extraction[J]. IEEE transactions on knowledge and data engineering, 1995, 7(5): 713－724.

[3] RILOFF E, SHOEN J. Automatically acquiring conceptual patterns without an annotated corpus[C]//Proceedings of the 3rd workshop on very large corpora. Cambridge: Massachusetts Institute of Technology, 1995: 148－161.

[4] 姜吉发. 一种事件信息抽取模式获取方法[J]. 计算机工程, 2005, 31(15):96－98.

[5] 许君宁,董萍,刘怀亮. 基于知网的中文事件抽取研究[J]. 情报杂志,2009,28(12):150－151,137.

[6] AHN D. The stages of event extraction [C]//Proceedings of the workshop on annotations and reasoning about time and events. Stroudsburg: Association for Computational Linguistics, 2006: 1－8.

[7] SAHA S, MAJUMDER A, HASANUZZAMAN M, et al. Bio-molecular event extraction using support vector machine [C]//Proceedings of the 3rd international conference on advanced computing. Piscataway: IEEE, 2011: 298－303.

[8] ZHU F, LIU Z T, YANG J L, et al. Chinese event place phrase recognition of emergency event using maximum entropy [C]//Proceedings of IEEE international conference on cloud computing and intelligence systems. Washington: IEEE Computer Society, 2011: 614－618.

[9] 许旭阳,李弼程,张先飞,等. 基于事件实例驱动的新闻文本事件抽取[J]. 计算机科学,2011,38(8):232－235.

[10] 刘振. 基于网络科技信息的事件抽取研究[J]. 情报科学,2018,36(9):115－117,122.

[11] 吉久明,陈锦辉,李楠,等. 中文事件抽取研究文献之算法效果分析[J]. 现代情报,2015,35(12):3－10.

[12] CHEN Y, XU L, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing. Stroudsburg: Association for Computational Linguistics, 2015: 167－176.

[13] NGUYEN T, GRISHMAN R. Modeling skip-grams for event detection with convulutional neural networks [C]//Proceedings of the 2016 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics,

2016: 886 – 891.
[14] FENG X C, QIN B, LIU T. A language-independent neural network for event detection [J]. Science China information sciences, 2018, 61(9): 66 – 71.
[15] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural network [C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2016: 300 – 309.
[16] ZHAO Y, JIN X, WANG Y, et al. Document embedding enhanced event detection with hierarchical and supervised attention [C]//Proceedings of the 56th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 414 – 419.
[17] HONG Y, ZHOU W, ZHANG J, et al. Self-regulation: employing a generative adversarial network to improve event detection [C]//Proceedings of the 56th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 515 – 526.
[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 annual conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171 – 4186.
[19] JIN Y L, XIE J F, GUO W S, et al. LSTM-CRF neural network with gated self attention for Chinese NER[J]. IEEE access, 2019, 4(4): 136694 – 136703.
[20] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer science, 2015, 5: 177 – 181.

作者贡献说明：
刘忠宝：确定研究选题，撰写论文；
党建飞：数据采集与实验设计；
张志剑：辅助文献收集，实验算法实现。

Research on Automatic Extraction of Historical Events and Construction of Event Graph Based on *Historical Records*

Liu Zhongbao[1,2] Dang Jianfei[2] Zhang Zhijian[2]
[1] Key Laboratory of Cloud Computing and Internet-of-Things Technology (Quanzhou University of Information Engineering), Fujian Province University, Quanzhou 362000
[2] School of Software, North University of China, Taiyuan 030051

Abstract: **[Purpose/significance]** *Historical Records* is the first biographical history book in China, which contains almost all the significant historical events during more than 3000 years between the Yellow Emperor and the Emperor Wu of Han. How to efficiently extract these historical events and their relationships is quite important to penetrate the historical appearances, reveal the historical essences and discover the historical laws. **[Method/process]** The BERT model and LSTM-CRF model were introduced in this paper, and historical events extraction method based on *Historical Records* was proposed and the historical event graph was constructed. **[Result/conclusion]** The experiment results show that the F1 values of historical event and its components extraction are respectively 0.823 and 0.760. The rare known knowledge is invented by the event graph, which providing essential literature foundation for many researchers, such as philology, history and sociology, to conduct their researches.

Keywords: *Historical Records* extraction of historical events event graph bidirectional encoder representations from transformers (BERT) bidirectional long short-term memory (BiLSTM) conditional random field (CRF)